



گزارش پروژه نهایی

شماره دانشجویی: ۹۸۱۰۰۱۱۸

نام و نام‌خانوادگی: پرهام چاوشیان

مقدمه

در این پروژه هدف بررسی داده‌های *microarray* برای بیماری *AML* بوده است. در پایان این گزارش به طور مفصل در مورد این بیماری توضیح داده شده است، اما به طور خلاصه این بیماری نوعی سرطان خون و مغز استخوان به حساب می‌آید که در آن سلول‌های *myeloid* بیش از حالت عادی تولید می‌شوند. این مشکل منجر به افزایش تولید پیش‌سازهای گلبول‌های سفید می‌شود که توانایی مقابله با عفونت را ندارند و بدین ترتیب بدن در برابر عفونت ضعیف می‌شود. این بیماری به کم‌خونی و کاهش تعداد گلبول‌های قرمز منجر می‌شود و در عوض آن سلول‌های خونی غیرعادی تولید می‌شوند که سبب نارسایی‌های گوناگون می‌شوند.

در این پروژه قصد داریم تا از مجموعه داده *GSE48558* استفاده کنیم و ارتباط بین تغییر بیان ژن‌ها در افراد مبتلا و افراد سالم را بیابیم.

تحلیل کد

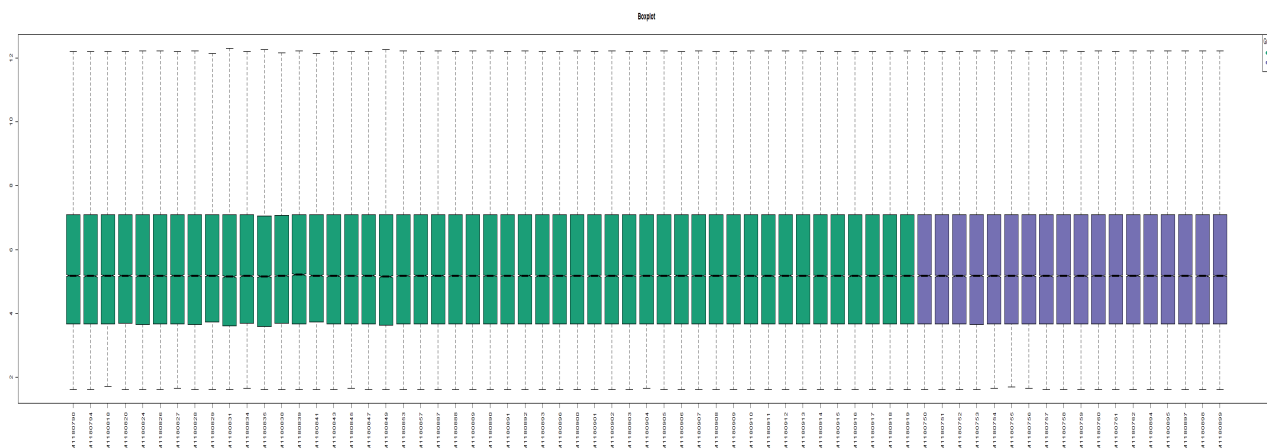
برای تحلیل این پروژه از زبان برنامه‌نویسی *R* استفاده شده است و تمامی کدها در محیط *RStudio* اجرا شده‌اند. برای اجرا کردن زبان *R*، از نسخه ۴.۱.۲ این زبان استفاده شده است. در ادامه عملکرد کلی هر بخش را توضیح خواهیم داد. ابتدای هر بخش در کد با کامنتی با همان عنوان مشخص شده است.

بخش *Libraries*: در این بخش تمامی کتابخانه‌ها مورد نیاز که برای بارگذاری داده‌ها، رسم نمودار، تحلیل، بررسی و محاسبه مقادیر و ... استفاده شده‌اند آورده شده است.

بخش *LoadData*: در این بخش با استفاده از کتابخانه *GEOquery* داده‌های مربوط به داده‌ها را بارگذاری می‌کنیم. سپس با استفاده از اطلاعات داده شده، داده‌ها را با برچسب‌های *T* به معنای سرطانی و *N* به معنای نرمال، برچسب‌گذاری می‌کنیم. این برچسب‌ها در متغیری به نام *group* ذخیره می‌شوند.

بخش *DeleteExtraSamples*: در این بخش داده‌هایی را که در هیچ‌کدام از دسته‌های تست و نرمال قرار ندارند را حذف می‌کنیم، البته این عمل اختیاری است. اگر بخواهید سایر داده‌ها را هم در بررسی داشته باشید، کافی است که این تکه از کد را کامنت کنید و برنامه را اجرا کنید.

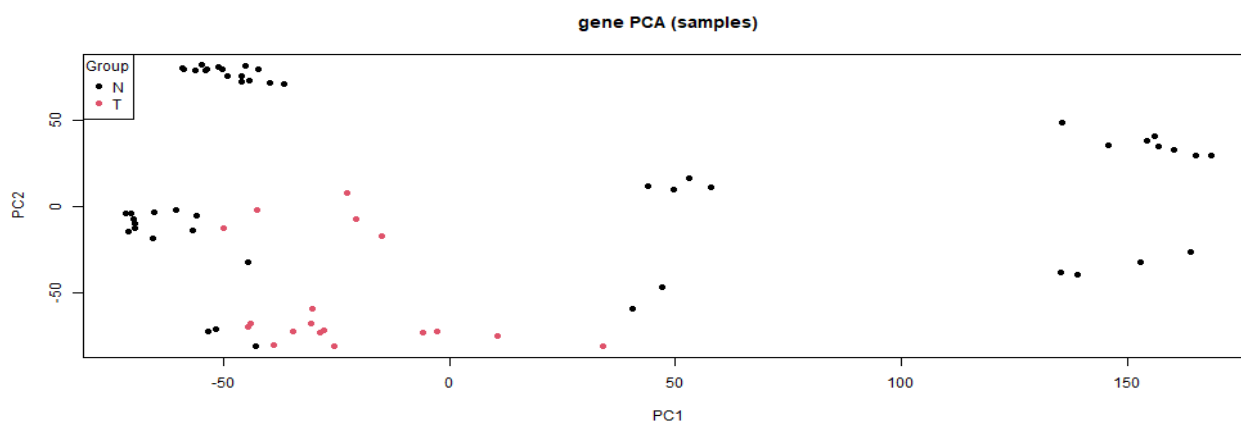
بخش *DrawBoxplot*: در این بخش نحوه توزیع ژن‌ها مختلف را بررسی می‌کنیم. ابتدا اطلاعات داده‌ها را در یک متغیر نگه می‌داریم و سپس اندیس برچسب‌های مرتب شده را به دست می‌آوریم. در نهایت با استفاده از توابع کتابخانه‌ای، نمودار زیر به دست می‌آید:



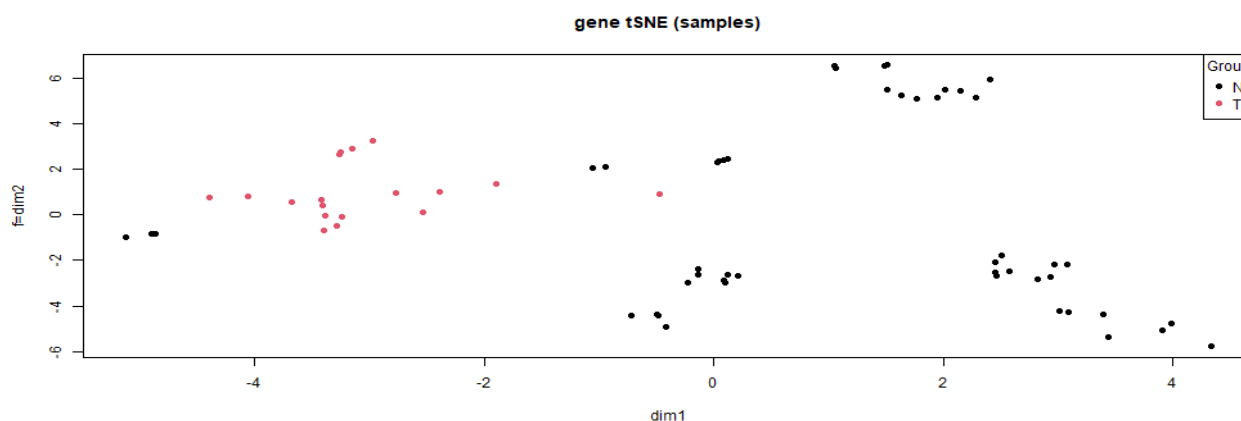
تحلیل و بررسی شکل داده شده نشان می‌دهد که در درجه اول داده‌ها در مقیاس لگاریتمی هستند و احتیاجی به تغییر مقیاس نداریم. همچنین نزدیک بودن میانه، چارک اول و چارک سوم توزیع‌ها، نشان می‌دهد که احتیاجی به نرمالیزه کردن

داده‌ها نداریم.

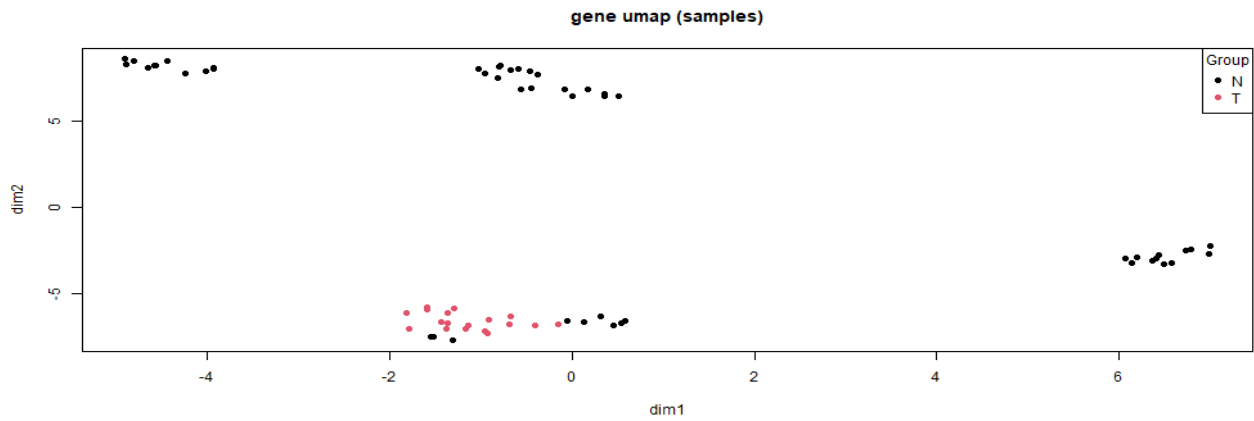
بخش *Dimension Reduction* در این بخش با استفاده از ۳ الگوریتم متفاوت بعد داده‌های ورودی را به دو بعد کاهش می‌دهیم تا قابل ترسیم و بررسی باشند. به ازای هر الگوریتم، کد دو خروجی تولید می‌کند. یکی از آن‌ها مربوط به کاهش بعد هر *sample* است (حالتی که ژن‌ها ویژگی باشند) و خروجی دیگر مربوط به کاهش بعد هر ژن است (حالتی که *sample*ها ویژگی باشند). خروجی دوم به صورت حاشیه‌ای تولید شده است و خروجی اصلی که ما مورد بررسی قرار می‌دهیم همان الگوریتم *PCA* : ابتدا نتیجه حاصل از این روش را می‌آوریم:



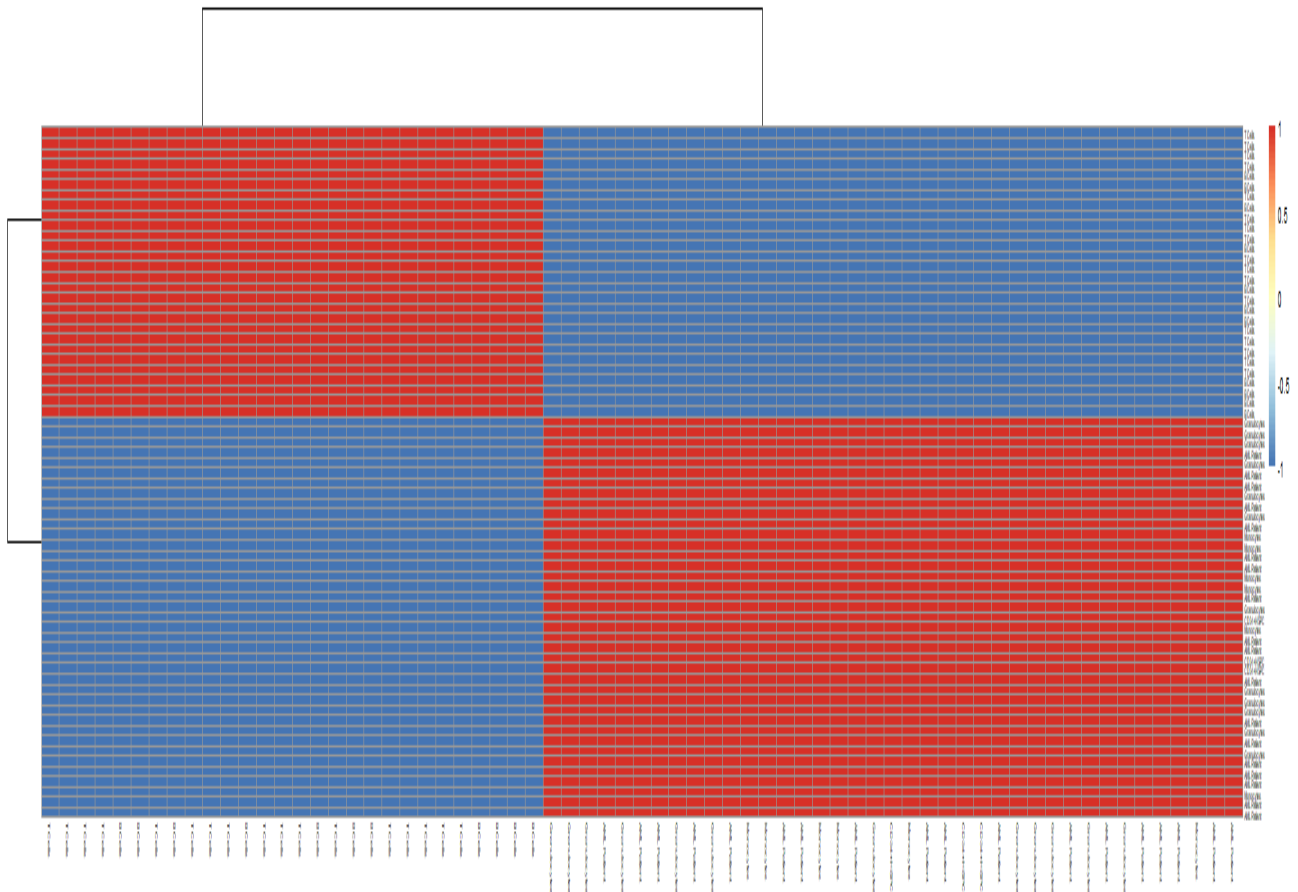
همانطور که در شکل مشاهده می‌شود، این الگوریتم توانسته تفکیک نسبتاً خوبی از *sample*ها نرمال و تست را نشان دهد. اما همانطور که قابل مشاهده است، نمونه‌ها نرمال، در دو خوشه توزیع شده‌اند، یکی از این خوشه‌ها، بسیار نزدیک به خوشه تست است. الگوریتم *tSNE* : نتیجه حاصل شده از این روش به شکل زیر است:

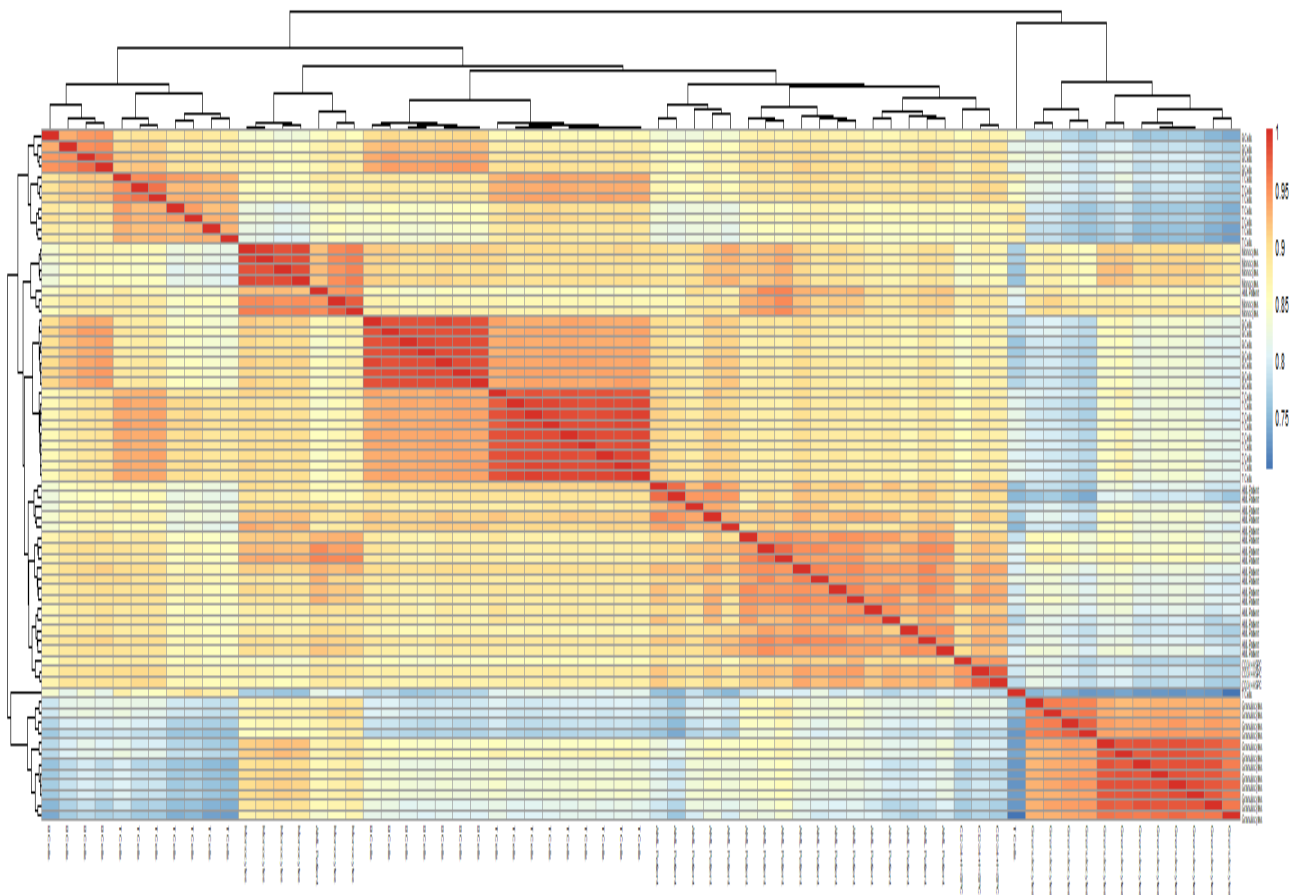


این الگوریتم در جداسازی داده‌ها نسبت به الگوریتم قبلی، عملکرد بهتری داشته است، اما مشاهده می‌شود که سه تا از نمونه‌های نرمال همچنان در نزدیکی این نمونه‌ها تست قرار گرفته‌اند و بالعکس؛ که می‌توان اینطور برداشت کرد که این نمونه‌ها در ابعاد بالا نیز نزدیک به نمونه‌های گروه مقابل شبیه‌اند. الگوریتم *UMAP* : نتیجه حاصل شده از این روش به شکل زیر است:



در این الگوریتم، فاصله‌ای که بین خوشه‌های تست و نرمال افتاده است، به نسبت دو الگوریتم قبلی بیشتر است، اما تعداد نمونه‌ها نرمال که در بین نمونه‌های تست قرار گرفته‌اند، از الگوریتم *tSNE* بیشتر است. بنا به توضیحات داده شده، ما در ادامه برای کاهش ابعاد، از الگوریتم *tSNE* استفاده می‌کنیم. بخش *Correlation* : در این بخش همسبستگی نمونه‌ها را با استفاده از دو *heatmap* بررسی می‌کنیم:





نمودار اول با استفاده از داده‌ها کاهش بعد یافته در قسمت قبل به دست آمده است، همانطور که مشاهده می‌شود، این نمودار اطلاعات مفید زیادی به ما نمی‌دهد. بنابراین تصمیم گرفتیم که از داده‌های اصلی نیز برای ساختن یک *heatmap* دیگر استفاده کنیم که حاصل آن شکل دوم شده است. همانطور که مشاهده می‌شود این شکل به خوبی روابط همبستگی بین نمونه‌ها را نشان می‌دهد و مقادیر آن صرفاً ۱ یا منفی ۱ نیستند. بر طبق این جدول بیشترین همبستگی منابع به تست، به ترتیب *monocytes* و *cd34* هستند.

بخش: *Differential Expression Analysis*: در ابتدای این بخش ماتریسی می‌سازیم که با استفاده از آن نمونه‌ها به صورت *OneHot* برچسب‌گذاری می‌شوند. از این ماتریس در ادامه استفاده خواهیم کرد. سپس با استفاده از توابع کتابخانه *limma* یک ماتریس که به نوعی بیانگر اختلاف بین نمونه‌ها با برچسب‌های متفاوت است به دست می‌آوریم و یک مدل خطی بر روی داده فیت می‌کنیم. برای اینکه بتوانیم از توزیع پیشین در مدل‌مان استفاده کنیم، یک مدل بیزی جدید با استفاده از مدل قبلی بر روی داده‌ها فیت می‌کنیم. در نهایت جدولی به دست می‌آید که شامل اطلاعات زیادی در مورد ژن‌ها است اما از بین تمامی اطلاعات ما تنها نام و نشانه ژن، *Adj P Value* و *logFC* را نگه می‌داریم. تفاوت معنادار را به این صورت تعیین کرده‌ایم که میزان بیان حداقل دو برابر یا حداکثر نصف شده باشد و از طرفی هم در بین ژن‌ها با این ویژگی، آن‌هایی که *Adj P Value* کمتر از ۵ درصد است را نگه می‌داریم. ژن‌های به دست آمده را در دو فایل ذخیره می‌کنیم. برخی از ژن‌ها ممکن است دارای چند نام باشند، نام‌های مختلف آن‌ها را در سطور مختلف می‌نویسیم.

آنالیز *gene ontology* و *pathway*:

ابتدا لازم به ذکر است که به علت کثرت ژن‌ها، ۱۰۰ ژن ابتدایی برای تحلیل به سایت *Enrichr* داده شده‌اند. ژن‌ها با بیان بالاتر: بر طبق خروجی سایت مذکور، بر اساس ۳ مجموعه *BioPlane 2019*، *KEG; 2021 Human* و *Elsevie; Pathway Collection* در میان *pathway*‌ها، *Cell cycle* بیشتر مورد توجه بوده است، که تصویر آن در ادامه آمده است.

- intracellular membrane-bounded organelle (GO:0043231)

حال به بررسی ژن‌ها با بیان پایین‌تر می‌پردازیم:
ژن‌ها با بیان پایین‌تر: برخلاف قسمت قبل، مسیری که اکثریت مجموعه‌ها در اولویت اول قرار دهند وجود نداشت. برای مجموعه BioPlanet 2019 مورد توجه‌ترین مسیر *Interleukin – 2 signaling pathway* بوده است و برای مجموعه KEGG 2021 Human مورد توجه‌ترین مسیر *Sphingolipi; signaling pathway* بوده است. مجدداً در زمینه تحلیل *ontology*‌ها، در سه زمینه زیر موارد قابل توجه را بیان می‌کنیم:
زمینه *Biological Process* 2021 : طبق *GO Biological Process* 2021 موارد زیر در این زمینه قابل توجه‌اند:

- positive regulation of type I interferon-mediated signaling pathway (GO:0060340)
- positive regulation of innate immune response (GO:0045089)

زمینه *Molecular Function* 2021 : طبق *GO Molecular Function* 2021 موارد زیر در این زمینه قابل توجه‌اند:

- TAP1 binding (GO:0046978)

زمینه *Cellular Component* 2021 : طبق *GO Cellular Component* 2021 موارد زیر در این زمینه قابل توجه‌اند:

- phagocytic vesicle membrane (GO:0030670)

توضیحاتی درباره AML

AML در واقع اختصاری برای *Acute Myeloid Leukemia*، بنابراین بهتر است پیش از آنکه درباره AML توضیح دهیم، بررسی کنیم که لوکمی چیست. به طور کلی، به تمامی سرطان‌هایی که در سلول‌هایی که قرار است به سلول خونی منجر شوند ایجاد می‌شود، لوکمی می‌گویند. رایج‌ترین شکل لوکمی در در حالات اولیه سلول‌های سفید خونی مشاهده می‌شود، اما این مسئله در سایر سلول‌ها نیز مشاهده شده است. لوکمی‌ها ممکن است در سلول‌های مغز استخوان یا لنفاوی ایجاد شوند. همچنین بر اساس سرعت رشد، لوکمی‌ها به دو دسته مزمن و حاد تقسیم می‌شوند.
AML لوکمی حاد مغز استخوان است، که در بافت مغز استخوان آغاز می‌شود. AML‌ها اکثراً به سرعت به خون منتقل می‌شوند و همچنین در موارد حادتر، به بخش‌های دیگر بدن مانند: مغز و نخاع، کبد و گره‌های لنفائی متاستاز می‌دهند. برخی عوامل افزایش احتمال ابتلا به AML: در اینجا چند مورد از مواردی که سبب افزایش احتمال ابتلا به AML می‌شوند را بررسی می‌کنیم:

- سن : احتمال ابتلا به این بیماری رابطه مستقیمی با سن فرد دارد و با افزایش سن، احتمال ابتلا نیز افزایش می‌یابد.
- جنسیت : طبق آمار، ابتلا به AML در مردان شایع‌تر از زنان است.
- استعمال دخانیات : استعمال دخانیات می‌تواند احتمال ابتلا به AML را (همانند بسیاری بیماری دیگر) افزایش دهد.
- مواد شیمیایی : قرار گرفتن در معرض برخی مواد شیمیایی مانند، مانند بنزن، می‌تواند احتمال ابتلا را افزایش دهد.
- مواد رادیواکتیو : قرار گرفتن در معرض پرتوهای رادیواکتیو می‌تواند ریسک ابتلا را افزایش دهد.
- سابقه خانوادگی : احتمال ابتلا در افرادی که ساقه این بیماری را در خانواده خود داشته‌اند بیشتر است.

علائم مبتلا شدن به بیماری AML :

این بیماری علائم مختلفی دارد که می‌توان آن‌ها را به چند دسته تقسیم کرد:
دسته علائم عمومی: علائم زیر در بین بیماران AML شایعند، البته به ذکر است که این علائم به AML منحصر نیستند و تنها نمی‌توان به همین علائم برای تشخیص اکتفا کرد:

- خستگی

- کاهش اشتها
- کاهش وزن
- تب
- عرق‌های شبانه

دسته علائم ناشی از افزایش سلول‌های لوکمی: سلول‌ها لوکمی به طور متوسط از سلول‌ها خونی نرمال بزرگترند. این موضوع منجر به این می‌شود که تعداد زیادی از آن‌ها سبب ایجاد لخته خونی و انسداد رگ بشوند. که ممکن است این موضوع به سبب هم منجر شود، در این صورت به این بیمار باید به طور اورژانسی رسیدگی شود. به این وضعیت *Leukostasis* گفته می‌شود که از جمله علائم آن به شرح زیر است:

- تکلم گنگ
- ضعف در یک سمت بدن
- گیجی
- سردرد
- خواب آلودگی

دسته علائم مربوط به کاهش سلول‌های خونی قرمز: کاهش این سلول‌ها در بدن علل مختلفی دارد، اما یکی از علل آن، ابتلا به *AML* است. از علائم آن می‌توان به موارد زیر اشاره کرد:

- خستگی
- ضعف
- سردرد
- سرگیجه
- تنگی نفس
- احساس لرز
- رنگ‌پریدگی

دسته علائم مربوط به کاهش سلول‌های خونی سفید نرمال: از آنجایی که این سلول‌ها مسئول محافظت از بدن در برابر بیماری‌ها و عفونت‌ها هستند، بنابراین در صورت کاهش آن‌ها، مشکلات عفونی در کوتاه‌مدت بهبود نمی‌یابند و فرد در فواصل زمانی کوتاهی دچار مشکل می‌شود. البته لازم به ذکر است که در صورت ابتلا به *AML* ممکن است نه تنها شخص دچار کاهش سلول سفید نشود، بلکه این سلول‌ها افزایش هم پیدا کنند، اما همگی آن‌ها غیر نرمال و در دفاع از بدن بی‌تاثیر هستند.

دسته علائم مربوط به کاهش پلاکت: کمبود پلاکت به شکل‌ها زیر می‌تواند خودش را بروز دهد:

- خونریزی‌ها شدید و مکرر بینی
- خونریزی لثه
- خونریزی‌های طولانی به طوری که خون بند نیاید
- کبودی در نواحی مختلف پوست

دسته علائم شرایط حاد بیماری و انتشار آن به سایر بخش‌ها بدن:

- صرع
- مشکل در تعادل
- تاری دید
- بی‌حسی صورت
- ضعف
- سردرد
- تهوع و استفراغ

علل بروز بیماری *AML*:

بع‌طور کلی در هر سلول، ژن‌هایی وجود دارند که مسئول تکثیر و مرگ سلولند. به ژن‌های مسئول رشد و تکثیر سلول، *Oncogene* و به ژن‌ها مسئول مرگ و میر سلول، *Tumor Suppressor Gene* گفته می‌شود. گاهی اوقات در فرایند تکثیر سلول، ممکن است این ژن‌ها دچار تغییر شوند به گونه‌ای که ژن‌های تکثیر بیش از حد فعال شوند یا ژن‌های مرگ غیرفعال شوند. در چنین شرایطی است که رشد و تکثیر سلول از حالت عادی خارج شده و به تومور و سرطان منجر می‌شود. تغییرات رخ داده ممکن است به شکل‌های جابه‌جایی، حذف یا معکوس شدن کروموزوم باشند.

برخی ژن‌های مهم در ابتلا به *AML*:

طبق تحقیقات از جمله ژن‌های مهم در بیماری *AML*، ژن‌های *FLT3* و *KIT* هستند که در میان ژن‌ها با بیان بالاتر در نتایج ما، به ترتیب در رتبه‌های ۲ و ۲۰ ظاهر شده‌اند. البته برخی ژن‌ها نیز مانند ژن *CEBPA* و *NPM1* که بنابر تحقیقات مهم خوانده شده‌اند اما در میان نتایج ما در رتبه‌های پایین قرار گرفته‌اند. برای اطلاعات تکمیلی می‌توانید به این لینک مراجعه کنید.

درمان بیماری *AML*:

یکی از داروهایی که در بزرگسالان برای درمان این بیماری استفاده می‌شود *XOSPATA* است. این دارو بر روی ژن *FLT3* که بالاتر در مورد آن صحبت شد اثرگذار است. برای اطلاعات بیشتر این لینک را مطالعه کنید. همچنین در این لینک می‌توانید لیستی از داروهایی که برای درمان *AML* کاربرد دارند را مشاهده کنید.